Pr3_Anàlisis_dades_titanic

Ricard Deza Tripiana

29 de desembre, 2018

Contents

1. Descripció del dataset

> *Perquè és important i quina pregunta/problema pretén respondre?*

El dataset que tractarem en aquesta pràctica contè les dades dels passatger del Titanic. Com ja sabem, el Titanic va ser una embarcació transatlàntica la qual va patir un accident i naufragà en el seu viatge de inauguració.

L'objectiu d'aquesta pràctica és la neteja, tractament i anàlisis de les dades per tal de poder respondre a la pregunta sobre si existeix algun grup de passatgers amb més probabilitats de sobreviure a l'accident.

En primer lloc, per tal de poder analitzar el conjunt de dades, llegirem els fitxers proporcionats. Disposem de dos fitxers, train i test, per tal de poder generar un model i poder provar-lo. A partir d'ara sempre parlarem del conjunt de dades d'entrenament. En el cas que es tracti del conjunt de prova, ho especificarem.

```r
# Lectura del fitxer en un dataframe (train)
passatgers <- read.csv(paste(ruta, "train.csv", sep = ""), header = TRUE, sep = ",",
    na.strings = "NA", encoding = "UTF-8")
# Lectura del fitxer en un dataframe (test)
passatgers_test <- read.csv(paste(ruta, "test.csv", sep = ""), header = TRUE, sep = ",",
    na.strings = "NA", encoding = "UTF-8")
# Primers registres del datatset
kable(head(passatgers[, 1:6]))
```

PassengerId	Survived	Pclass	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22

PassengerId	Survived	Pclass	Name	Sex	Age
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
3	1	3	Heikkinen, Miss. Laina	female	26
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
5	0	3	Allen, Mr. William Henry	male	35
6	0	3	Moran, Mr. James	male	NA

```r
head(passatgers[, 7:12])
```

SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	A/5 21171	7.2500		S
1	0	PC 17599	71.2833	C85	C
0	0	STON/O2. 3101282	7.9250		S
1	0	113803	53.1000	C123	S
0	0	373450	8.0500		S
0	0	330877	8.4583		Q

El dataset conté les dades de 891 passatgers i 12 variables per cadascún d'ells:

```r
# Número de passatgers
nrow(passatgers)
## [1] 891
# Número de variables
ncol(passatgers)
## [1] 12
```

Aquestes variables són les següents:

```r
# Nom de les variables
labels(passatgers)[2]
## [[1]]
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

La primera variable, *PassangerId*, tan sols identifica el passatger amb un identificador numèric seqüencial.
La variable *Survived* indica si el passatger va sobreviure al accident, indicant un 1 si va sobreviure o un 0 en el cas contrari.
La variable *Pclass* indica la classe socio-econòmica del passatger. En aquest cas tenim els valors 1, 2 i 3 corresponents a primera, segona i tercera classe, respectivament.
La variable *Name* conté el nom dels passatger.
La variable *Sex* indica el sexe del passatger. Observant les dades veiem que es descriu amb **male** als homes i **female** a les dones.
La variable *Age* indica l'edat en anys del passatger. En el cas que l'edat sigui menor a 1 any, aquesta serà fraccionada. En el cas que sigui estimada, serà del tipus xx.5.
La variable *SibSp* indica el nombre de germans o cònjugues a bord del Titanic.
La variable *Parch* indica el nombre de pares o fills a bord del Titanic. En el cas que un menor viatgés acompanyat de una cuidadora, el valor d'aquesta variable és 0.
La variable *Ticket* indica el codi del ticket d'embarcament.
La variable *Fare* indica la tarifa que va pagar el passatger.
La variable *Cabin* indica el codi del camarot que ocupava el passatger en l'embarcació.
Finalment, la variable *Embarked* indica el port en el qual pujà a bord del vaixell. Els valors són: **C** = Cherbourg, **Q** = Queenstown i **S** = Southampton.

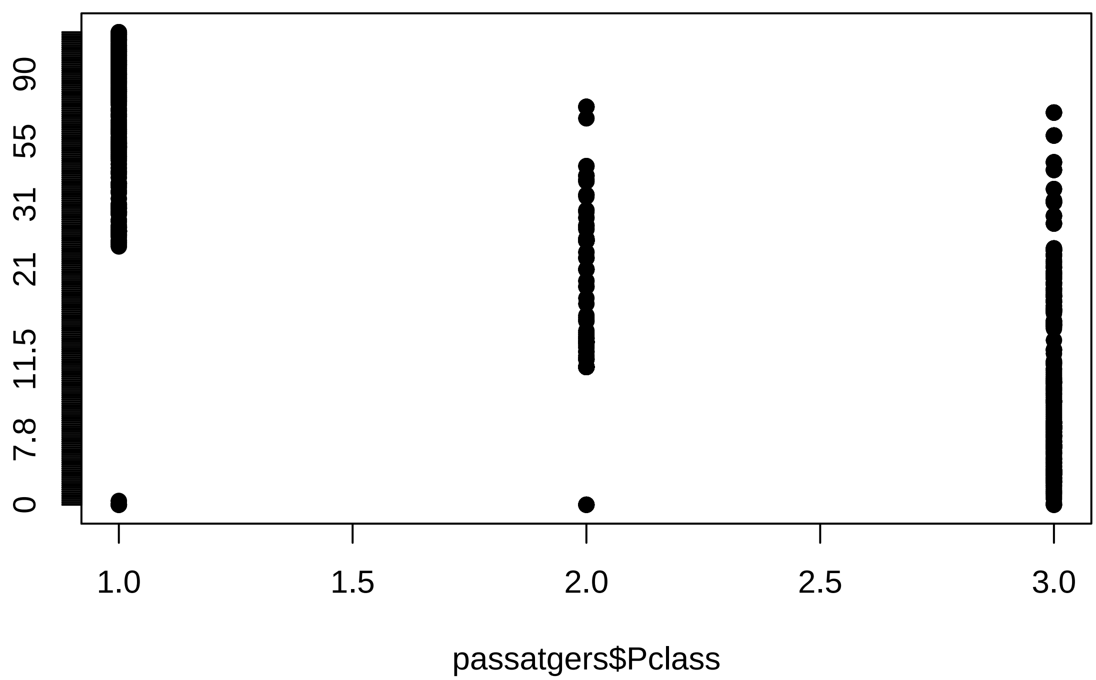
2. Integració i selecció de les dades d'interès a analitzar

Per tal de realitzar l'anàlisis, ens quedarem amb les variables d'interès per tal de predir la probabilitat de sobreviure. Considerem que aquestes variables seran:

- Survived
- Pclass
- Sex
- Age
- SibSp
- Parch

La resta de variables considerem que no son rellevants per tal d'establir una predicció de supervivència. En el cas de les variables *PassangerId*, *Name*, *Ticket*, *Embarked* són identificadors que entenem que no afectaran en la probabilitat de sobreviure.
En el cas de la variable *Cabin*, si realitzem un petit anàlisis observem que conté molts valors perduts.
Per últim, en el cas de la variable *Fare* veiem una multicol · linealitat amb la variable *Pclass*:

```
# Mostrem gràfica comparant Pclass i Fare
stripchart(passatgers$Pclass ~ passatgers$Fare, pch = 19)
```



Com s'observa en el gràfic, no s'inclou la variable *Fare*, degut que existeix una dependència entre les variables *Pclass* i *Fare* Això implica un problema de multicol · linealitat, la qual cosa ocasiona efectes molt importants en les estimacions i els resultats poden ser confusos.

3. Neteja de les dades

Abans de netejar les dades, procedirem a realitzar una primera revisió d'aquestes.

La variable *Survived* hauria d'estar factoritzada amb els valors 0 i 1. Si realitzem un resum de la variable, obtenim:

```
# Resum de la variable Survived
summary(passatgers$Survived)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3838  1.0000  1.0000
```

Com veiem, no s'està detectant la variable com un factor, per tant, l'haurem de factoritzar.

```
# Factoritzem la variable Survived
passatgers$Survived <- as.factor(passatgers$Survived)
is.factor(passatgers$Survived)
## [1] TRUE
```

Si realitzem un resum de la variable ara, observem com tenim 342 supervivents i 549 no supervivents:

```
# Resum de la variable Survived
summary(passatgers$Survived)
##   0   1
## 549 342
```

En el cas de la variable *Pclass*, ens passa el mateix. La variable hauria d'estar factoritzada amb els valors 1, 2 i 3. Si realitzem un resum de la variable, obtenim:

```
# Resum de la variable Pclass
summary(passatgers$Pclass)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.309   3.000   3.000
```

Com veiem, no s'està detectant la variable com un factor, per tant, l'haurem de factoritzar.

```
# Factoritzem la variable Pclass
passatgers$Pclass <- as.factor(passatgers$Pclass)
passatgers_test$Pclass <- as.factor(passatgers_test$Pclass)
is.factor(passatgers$Pclass)
## [1] TRUE
```

Si realitzem un resum de la variable ara, observem com tenim 216 passatgers de primera classe, 184 de segona classe i 491 de tercera classe:

```
# Resum de la variable Pclass
summary(passatgers$Pclass)
##   1   2   3
## 216 184 491
```

Si observem la variable *Sex*, observem com tenim 314 dones i 577 homes:

```
# Resum de la variable Sex
summary(passatgers$Sex)
## female   male
##    314    577
```

En aquest cas no hem de realitzar cap tractament d'inici.

Si analitzem la variable *Age*, observem el següent resum:

```
# Resum de la variable Age
summary(passatgers$Age)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

Veiem que apareixen molts valors NA. Aquesta casuística la tractarem més endavant.

Respecte la variable *SibSp*, observem el següent:

```
# Resum de la variable SibSp
summary(passatgers$SibSp)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.523   1.000   8.000
```

En aquest cas, no veiem cap valor NA, però veiem un valor que s'allunya molt de la mitjana i observem zeros. Aquesta casuística també la tractarem més endavant.

Per últim, analitzem la variable *Parch*:

```
# Resum de la variable Parch
summary(passatgers$Parch)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3816  0.0000  6.0000
```

Com en el cas anterior, no veiem cap valor NA, però veiem un valor que s'allunya molt de la mitjana i observem zeros.

3.1 Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Hem observat que tenim variables amb elements buits (NA) o que contenen zeros. Procedim a gestionar aquestes casuístiques.

Pel que fa a la variable *Age*, hem vist que conté 177 registres sense l'edat informada. Considerarem que no va nèixer cap nen o nena durant el viatge. Entenem que una edat 0 no pot ser, per tant, per solucionar aquest escenari, imputarem els valors perduts utilitzant el mètode dels k-veïns més propers usant la distància de Gower.

```
# Mètode KNN per imputar valors perduts
passatgers_2 <- kNN(passatgers)[, 1:12]
passatgers_test <- kNN(passatgers_test)[, 1:12]
summary(passatgers_2$Age)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   21.00   28.00   29.45   36.75   80.00
```

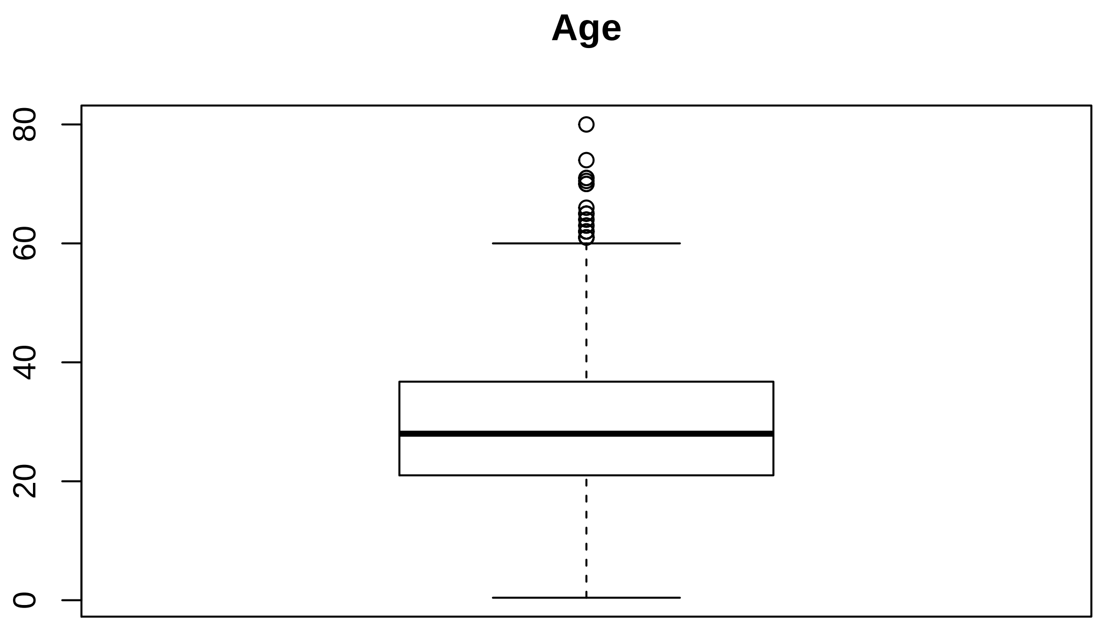
Podem observar com ja no tenim cap passatger amb el valor de la variable *Age* buit.

Pel que fa a les variables *SibSp* i *Parch*, contenen valors zeros. En el cas de la variable *SibSp* és una casuística plausible degut que un passatger podria viatgar sense germans ni cònjugues (fill únic i solter, o simplement viatja sol). En el cas de la variabl *Parch* abans hem comentat que si un menor viatjava acompanyat de una cuidadora, el valor d'aquesta variable és 0.

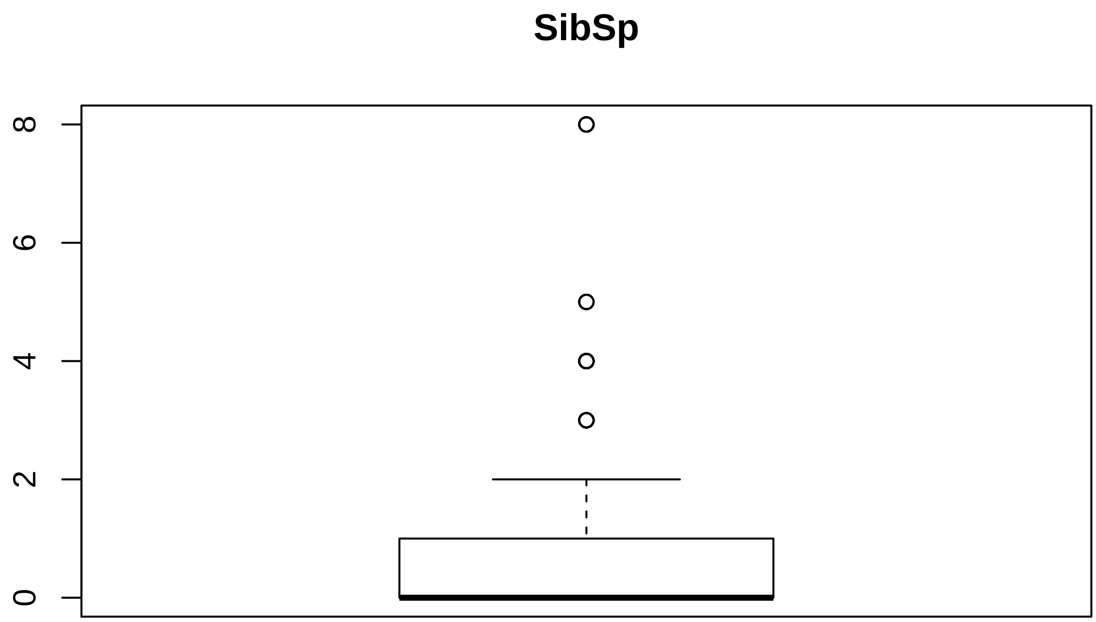
3.2 Identificació i tractament de valors extrems

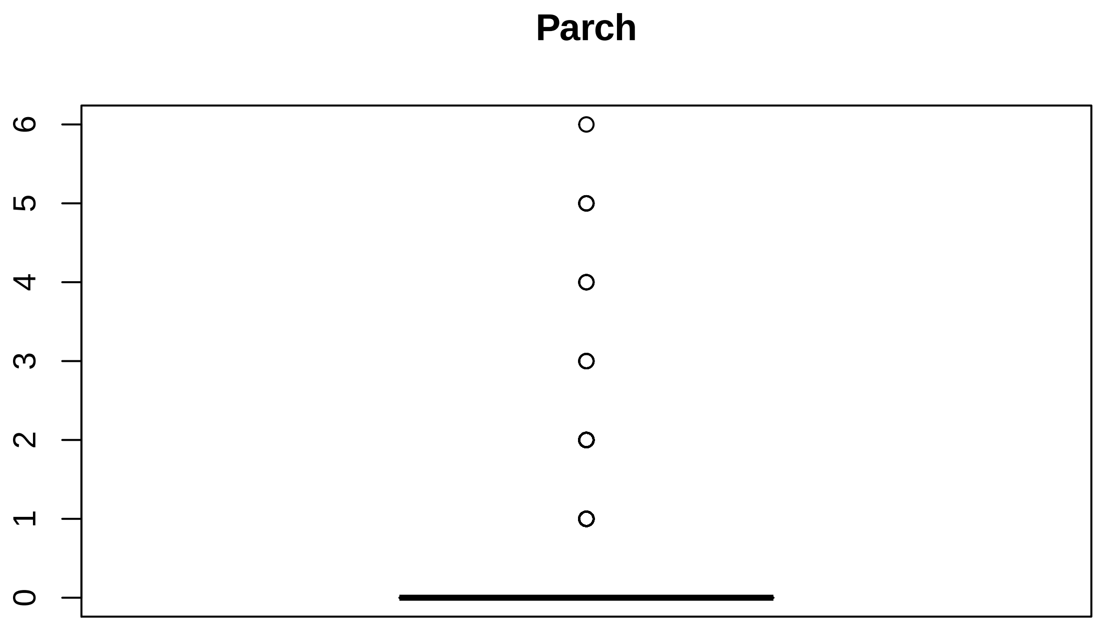
Per tal d'identificar possibles valors extrems, mostrarem gràfics de caixa per les variables *Age*, *SibSp* i *Parch*:

```
# Boxplots Age, SibSp i Parch
variables <- c("Age", "SibSp", "Parch")
for (name in variables) {
    boxplot(passatgers_2[, name], main = name)
}
```



Age
0
20
40
60
80


SibSp
8
6
4
2
0



Els extrems que es mostren en els gràfics són el següents:

```
# Valors extrems
boxplot.stats(passatgers_2$Age)$out
##  [1] 66.0 65.0 71.0 70.5 61.0 61.0 62.0 63.0 65.0 61.0 64.0 65.0 63.0 71.0
## [15] 64.0 62.0 62.0 61.0 61.0 80.0 70.0 70.0 62.0 74.0
boxplot.stats(passatgers_2$SibSp)$out
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3
## [36] 5 4 3 4 8 4 3 4 8 4 8
boxplot.stats(passatgers_2$Parch)$out
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1
##  [36] 2 1 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1
##  [71] 1 2 1 2 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2
## [106] 2 3 4 1 2 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1
## [141] 2 1 1 2 5 2 1 1 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1
## [176] 1 2 1 2 3 1 2 1 2 2 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1
## [211] 1 5 2
```

Podem observar que el valors extrems de la variable *Age* entren dins de la realitat.
En el cas de la variable *SibSp*, veiem valors extrems de 3, 4, 5 i 8 germans i cònjugues. Els tres primers casos, són molt probables, sabent que són dades de gent de principi del segle XX. En el cas del valor 8, tenim els següents passatgers:

```
# Passatgers amb 8 germans i cònjugues
passatgers[which(passatgers$SibSp == 8), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age

	PassengerId	Survived	Pclass	Name	Sex	Age
160	160	0	3	Sage, Master. Thomas Henry	male	NA
181	181	0	3	Sage, Miss. Constance Gladys	female	NA
202	202	0	3	Sage, Mr. Frederick	male	NA
325	325	0	3	Sage, Mr. George John Jr	male	NA
793	793	0	3	Sage, Miss. Stella Anna	female	NA
847	847	0	3	Sage, Mr. Douglas Bullen	male	NA
864	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NA

```
passatgers[which(passatgers$SibSp == 8), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
160	8	2	CA. 2343	69.55		S
181	8	2	CA. 2343	69.55		S
202	8	2	CA. 2343	69.55		S
325	8	2	CA. 2343	69.55		S
793	8	2	CA. 2343	69.55		S
847	8	2	CA. 2343	69.55		S
864	8	2	CA. 2343	69.55		S

```
# Passatgers amb 8 germans i cònjugues (test)
passatgers_test[which(passatgers_test$SibSp == 8), 1:6]
```

	PassengerId	Pclass	Name	Sex	Age	SibSp
189	1080	3	Sage, Miss. Ada	female	14.5	8
361	1252	3	Sage, Master. William Henry	male	14.5	8

```
passatgers_test[which(passatgers_test$SibSp == 8), 7:12]
```

	Parch	Ticket	Fare	Cabin	Embarked	PassengerId_imp
189	2	CA. 2343	69.55		S	FALSE
361	2	CA. 2343	69.55		S	FALSE

Podem observar que es tracta de 9 germans, és a dir, en principi la dada *SibSp* és correcta.

Si observem el valor extrem 5 per la variable *SibSp* i fem les mateixes comprovacions, veurem que el valor també és és correcte:

```
# Passatgers amb 5 germans i cònjugues
passatgers[which(passatgers$SibSp == 5), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
60	60	0	3	Goodwin, Master. William Frederick	male	11
72	72	0	3	Goodwin, Miss. Lillian Amy	female	16
387	387	0	3	Goodwin, Master. Sidney Leonard	male	1

	PassengerId	Survived	Pclass	Name	Sex	Age
481	481	0	3	Goodwin, Master. Harold Victor	male	9
684	684	0	3	Goodwin, Mr. Charles Edward	male	14

```
passatgers[which(passatgers$SibSp == 5), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
60	5	2	CA 2144	46.9		S
72	5	2	CA 2144	46.9		S
387	5	2	CA 2144	46.9		S
481	5	2	CA 2144	46.9		S
684	5	2	CA 2144	46.9		S

```
# Passatgers amb 5 germans i cònjugues (test)
passatgers_test[which(passatgers_test$SibSp == 5), 1:6]
```

	PassengerId	Pclass	Name	Sex	Age	SibSp
141	1032	3	Goodwin, Miss. Jessie Allis	female	10	5

```
passatgers_test[which(passatgers_test$SibSp == 5), 7:11]
```

	Parch	Ticket	Fare	Cabin	Embarked
141	2	CA 2144	46.9		S

Si observesim els altres valor de *SibSp*, ens trobaríem que els valors també són correctes.

Pel que fa a la variable *Parch*, ens surten valors extrems igual a 1. Això vol dir que la gran màjoria de passatgers viatjaven sense pares o fills, o bé menors acompanyats per una cuidadora. Si realitzem l'anàlisis sense tenir en compte aquests valors veiem el següent:

```
# Valors extrems sense 0
boxplot.stats(passatgers_2$Parch[which(passatgers_2$Parch != 0)])$out
##  [1] 5 5 4 4 4 4 5 5 6 5
```

Ara veiem que els valors extrems són 4, 5 i 6. Fem un anàlisis del valor 6:

```
# Passatgers amb 6 pares o fills
passatgers[which(passatgers$Parch == 6), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
679	679	0	3	Goodwin, Mrs. Frederick (Augusta Tyler)	female	43

```
passatgers[which(passatgers$Parch == 6), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
679	1	6	CA 2144	46.9		S

```
# Passatgers amb 6 pares o fills (test)
passatgers_test[which(passatgers_test$Parch == 6), 1:6]
```

	PassengerId	Pclass	Name	Sex	Age	SibSp
140	1031	3	Goodwin, Mr. Charles Frederick	male	40	1

```
passatgers_test[which(passatgers_test$Parch == 6), 7:12]
```

	Parch	Ticket	Fare	Cabin	Embarked	PassengerId_imp
140	6	CA 2144	46.9		S	FALSE

Observem que corresponen als pares dels germans 'Goodwin'. El que ens fa pensar és que dels pares dels 9 germans 'Sage' no tenim les dades.

Si observesim els altres valor de *Parch*, ens trobaríem que els valors també són correctes.

4. Anàlisi de les dades

L'objectiu que ens hem plantjat és estimar un model de regressió logística amb variable dependent *Survived* i com regressors, Pclass, Sex, Age, SibSp i Parch.

4.1 Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

En primer lloc, recodifiquem les variables *Pclass* i *Sex* de la següent forma:

```
library(car)
passatgers_2$Pclass <- recode(passatgers_2$Pclass, "1='H';2='M';3='L'")
summary(passatgers_2$Pclass)
##   H   L   M
## 216 491 184
passatgers_2$Sex <- recode(passatgers_2$Sex, "'male'='M';'female'='F'")
summary(passatgers_2$Sex)
##   F   M
## 314 577
```

A continuació creem les variables noves *PclassR* i *SexR* reordenades amb els valors de referència **H** i **F** respectivament:

```
# Reordenem la variable Pclass en la nova variable PclassR
passatgers_2$PclassR <- relevel(passatgers_2$Pclass, ref = "H")
head(passatgers_2$PclassR)
## [1] L H L H L L
## Levels: H L M
# Reordenem la variable Sex en la nova variable SexR
passatgers_2$SexR <- relevel(passatgers_2$Sex, ref = "F")
head(passatgers_2$SexR)
## [1] M F F F M M
## Levels: F M
```

4.2 Comprovació de la normalitat i homogeneïtat de la variància.

Si apliquem el model i realitzem la prova de homogeneïtat de variancies amb un nivell de significància del 0,05, obtenim el següent:

```
# Generem el model
model_logit <- glm(passatgers_2$Survived ~ passatgers_2$PclassR + passatgers_2$SexR +
    passatgers_2$Age + passatgers_2$SibSp + passatgers_2$Parch, family = "binomial")
summary(model_logit)
##
## Call:
## glm(formula = passatgers_2$Survived ~ passatgers_2$PclassR +
##     passatgers_2$SexR + passatgers_2$Age + passatgers_2$SibSp +
##     passatgers_2$Parch, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8253  -0.5914  -0.3911   0.6121   2.5077
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.653071   0.437068  10.646  < 2e-16 ***
## passatgers_2$PclassRL -2.678091   0.264324 -10.132  < 2e-16 ***
## passatgers_2$PclassRM -1.397327   0.271917  -5.139 2.77e-07 ***
## passatgers_2$SexRM    -2.752639   0.201310 -13.674  < 2e-16 ***
## passatgers_2$Age      -0.051613   0.008091  -6.379 1.78e-10 ***
## passatgers_2$SibSp    -0.421033   0.110779  -3.801 0.000144 ***
## passatgers_2$Parch    -0.078173   0.117778  -0.664 0.506862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  773.15  on 884  degrees of freedom
## AIC: 787.15
##
## Number of Fisher Scoring iterations: 5
```

Veiem que tots els regressors són significants, excepte *Parch*.

4.3 Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

Com que hem observat que la variable *Parch*, no és significant, crearem un model prescindint d'aquesta variables:

```
# Generem el model sense Parch
model_logit_2 <- glm(passatgers_2$Survived ~ passatgers_2$PclassR + passatgers_2$SexR +
    passatgers_2$Age + passatgers_2$SibSp, family = "binomial")
summary(model_logit_2)
##
## Call:
```

```
## glm(formula = passatgers_2$Survived ~ passatgers_2$PclassR +
##     passatgers_2$SexR + passatgers_2$Age + passatgers_2$SibSp,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8575  -0.5878  -0.3892   0.6185   2.5107
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            4.611676   0.432660  10.659  < 2e-16 ***
## passatgers_2$PclassRL -2.678956   0.264607 -10.124  < 2e-16 ***
## passatgers_2$PclassRM -1.397590   0.272046  -5.137 2.79e-07 ***
## passatgers_2$SexRM    -2.726827   0.197053 -13.838  < 2e-16 ***
## passatgers_2$Age      -0.051425   0.008083  -6.362 1.99e-10 ***
## passatgers_2$SibSp    -0.443313   0.106229  -4.173 3.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  773.6  on 885  degrees of freedom
## AIC: 785.6
##
## Number of Fisher Scoring iterations: 5
```

Per tal de escollir el millor model, extraiem l'AIC de cadascun d'ells i els comparem entre si per tal de triar el menor:

```
AIC(model_logit)
## [1] 787.1544
AIC(model_logit_2)
## [1] 785.6019
```

Com podem observar, el segon model sense la variable *Parch* és millor que el primer model calculat.

Amb aquest model, els coeficients de regressió són els següents:

```
# Coeficients de regressió
coeff_beta <- coefficients(model_logit_2)
coeff_beta
##           (Intercept) passatgers_2$PclassRL passatgers_2$PclassRM
##            4.61167591           -2.67895613           -1.39758959
##    passatgers_2$SexRM      passatgers_2$Age    passatgers_2$SibSp
##           -2.72682653           -0.05142501           -0.44331278
```

Per tant el model de regressió logística és:

$$Prob(Y_i = 1) = \frac{exp(\beta_0 + \beta_1 PclassRM_i + \beta_2 PclassRL_i + \beta_3 SexR_i + \beta_4 Age_i + \beta_5 SibSp_i)}{1 + exp(\beta_0 + \beta_1 PclassRM_i + \beta_2 PclassRL_i + \beta_3 SexR_i + \beta_4 Age_i + \beta_5 SibSp_i)}$$

Segons els valors obtinguts podem realitzar els anàlisis següents:

- El coeficient d'intersecció no té sentit analitzar-lo degut que la variable *Age* no pot ser 0.
- Si l'individu pertany a la classe 'H' (primera classe), la probabilitat de sobreviure depèn de la resta de variables *Age*, *SexR* i *SibSp*.

- Si l'individu pertany a la classe 'M' (segona classe), la probabilitat de sobreviure disminueix. De la mateixa forma, si pertany a la classes 'L' (tercera classe), la probabilitat de sobreviure disminueix encara més.
- Si l'individu és dona, la probabilitat de sobreviure depèn de la resta de variables *PclassRL*, *PclassRM*, *Age* i *SibSp*.
- Si l'individu és home, la probabilitat de sobreviure disminueix.
- Com més edat tingui l'individu, menys probabilitat de sobreviure té.
- Com més germans i/o cónjugues tingués a bord, menys probabilitat de sobreviure té.

Passem a anlitzar la qualitat d'ajust del model creat. En primer lloc, creem un dataframe on la primera columna sigui les observacións dels nostre conjunt de dades si un individu sobreviu o no, és a dir, la variable *Survived*, i la segona columna sigui els valors predits pel model anterior amb un llindar de discriminació del 70%:

```
# Calculem els valors predits
valors_predits <- predict(model_logit_2, passatgers_2, type = "response")
head(valors_predits)
##          1          2          3          4          5          6
## 0.08558537 0.90151858 0.64466494 0.91439256 0.06952673 0.13307588
# Interpretem els resultats amb el llindar indicat
clase_predita <- ifelse(valors_predits > 0.7, 1, 0)
head(clase_predita)
## 1 2 3 4 5 6
## 0 1 0 1 0 0
# Montem el data set a analitzar
data <- data.frame(obs = passatgers_2$Survived, pre = clase_predita)
kable(data.frame(Observació = head(data$obs), Predicció = head(data$pre)), align = c("l",
    "l"))
```

Observació	Predicció
0	0
1	1
1	0
1	1
0	0
0	0

5. Representació dels resultats a partir de taules i gràfiques.

Amb aquests resultats obtingut, podem montar una taula de doble entrada amb cadascuna de les variables anterior i obtenim la matriu de confusió.

```
# Montem la matriu de confusió
matriu_confusio <- table(data$obs, data$pre, dnn = c("Observació", "Predicció"))
matriu_confusio
##           Predicció
## Observació   0   1
##          0 530  19
##          1 162 180
```

Podem veure com tenim 19 falsos positius i 162 falsos negatius. Els falsos positius són aquells, en el nostre cas, que no sobreviuen i el model ha predit que si. En contra, els falsos negatius són aquells que sobreviuen i el model a predit que no.

En les següents taules mostrem els valors interessants que es poden extreure de la matriu de confusió:

```r
# Valors descriptius de la predicció
positius <- sum(data$obs == 1)
negatius <- sum(data$obs == 0)
positius_predit <- sum(data$pre == 1)
negatius_predit <- sum(data$pre == 0)
total <- nrow(data)
kable(data.frame(Mesura = c("Positius", "Negatius", "Positius predits", "Negatius Predits"),
    Valor = c(positius, negatius, positius_predit, negatius_predit)), align = c("l",
    "l", "l", "l"))
```

Mesura	Valor
Positius	342
Negatius	549
Positius predits	199
Negatius Predits	692

```r
tp <- sum(data$obs == 1 & data$pre == 1)
tn <- sum(data$obs == 0 & data$pre == 0)
fp <- sum(data$obs == 0 & data$pre == 1)
fn <- sum(data$obs == 1 & data$pre == 0)
kable(data.frame(Mesura = c("Certs positius", "Certs negatius", "Falsos positius",
    "Falsos negatius"), Valor = c(tp, tn, fp, fn)), align = c("l", "l", "l", "l"))
```

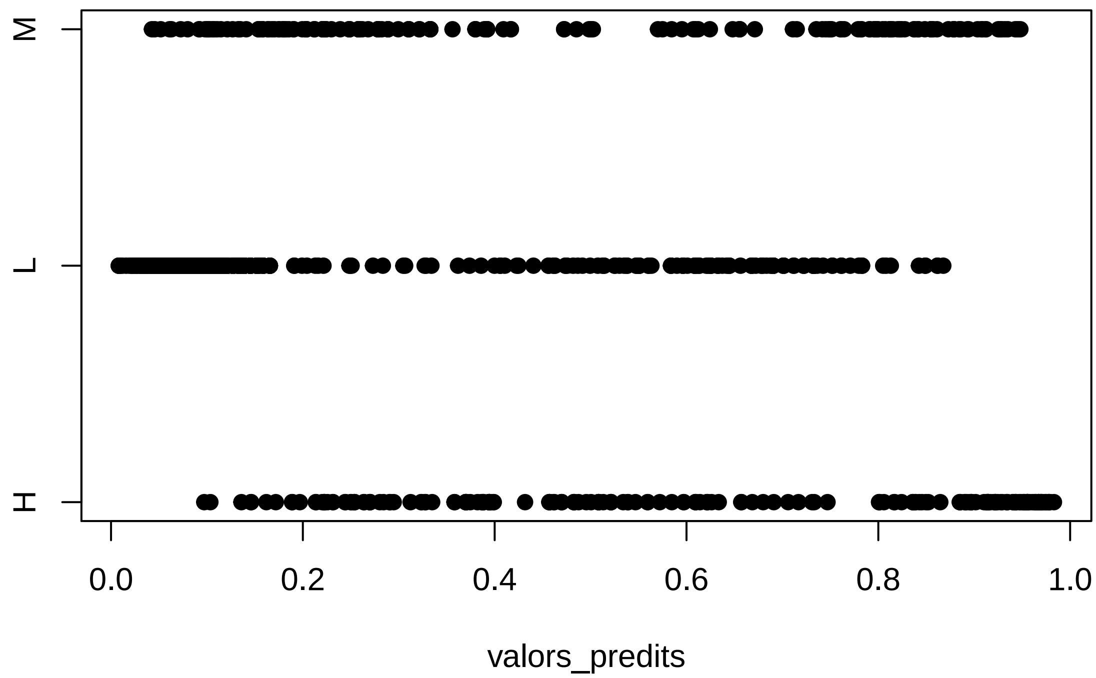
Mesura	Valor
Certs positius	180
Certs negatius	530
Falsos positius	19
Falsos negatius	162

```r
exactitut <- (tp + tn)/total
ratio_error <- (fp + fn)/total
sensibilitat <- tp/positius
especificitat <- tn/negatius
precisio <- tp/positius_predit
valor_pre_neg <- tn/negatius_predit
kable(data.frame(Mesura = c("Exactitud", "Ratio d'error", "Sensibilitat", "Especificitat",
    "Precisió", "Valor de predicció de negatius"), Valor = c(exactitut, ratio_error,
    sensibilitat, especificitat, precisio, valor_pre_neg)), align = c("l", "l", "l",
    "l", "l", "l"))
```

Mesura	Valor
Exactitud	0.7968575
Ratio d'error	0.2031425
Sensibilitat	0.5263158
Especificitat	0.9653916
Precisió	0.9045226
Valor de predicció de negatius	0.7658960

Ara mostrem un gràfic amb la comparació de la probabilitat de sobreviure amb la classe dels individus:

```
stripchart(valors_predits ~ passatgers_2$PclassR, pch = 19)
```



Com podem veure en aquest gràfic, hi ha una lleugera tendència a augmentar la probabilitat de sobreviure quan la classe és superior. Però observem casos classes **L** amb una probabilitat alta de sobreviure.
Si extreiem els individus de classe baixa amb una probabilitat superior al 70% de sobreviure, trobem el següent:

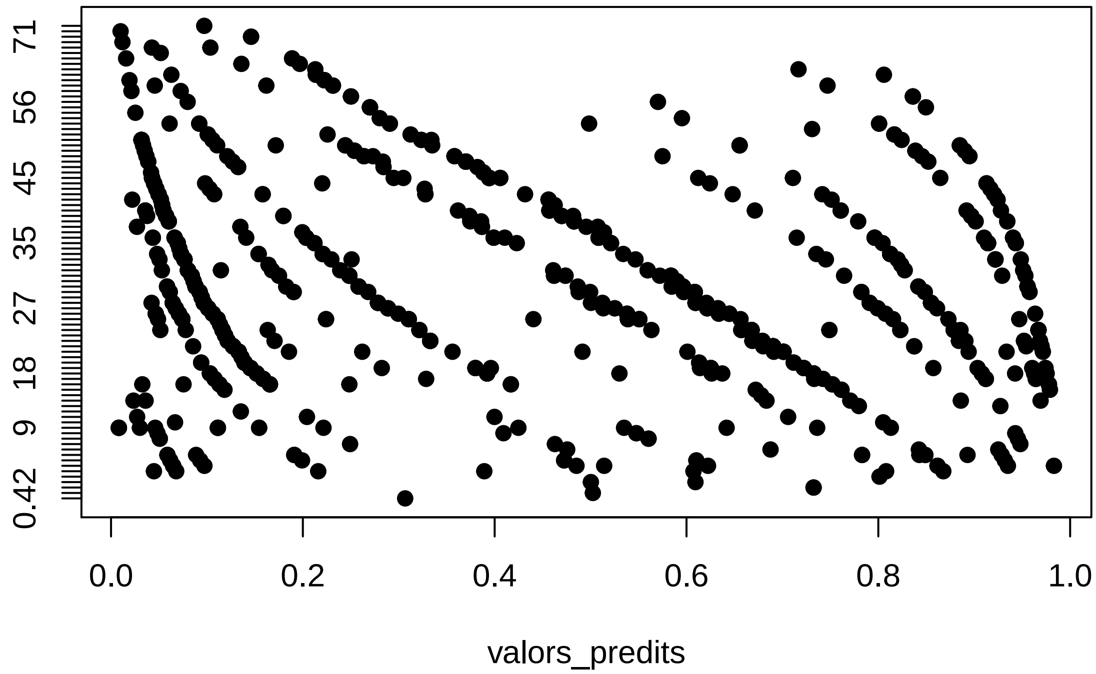
```
passatgers_2$Ppre <- valors_predits
passatgers_2$Pre <- data$pre
p_low_sur_pre <- passatgers_2[which(passatgers_2$Ppre >= 0.7 & passatgers_2$PclassR ==
    "L"), ]
kable(head(p_low_sur_pre[4:7]))
```

	Name	Sex	Age	SibSp
11	Sandstrom, Miss. Marguerite Rut	F	4	1
15	Vestrom, Miss. Hulda Amanda Adolfina	F	14	0
20	Masselmani, Mrs. Fatima	F	19	0
23	McGowan, Miss. Anna “Annie”	F	15	0
45	Devaney, Miss. Margaret Delia	F	19	0
107	Salkjelsvik, Miss. Anna Kristine	F	21	0

Podem observar que tots els individus són dones amb una edat menor a 21 anys. Com hem dit, el fet de ser dona fa pujar la probabilitat de sobreviure, com també l’augmenta tenir una edat baixa.

Si mostrem un gràfic amb la comparació de la probabilitat de sobreviure amb la edat dels individus:

```r
stripchart(valors_predits ~ passatgers_2$Age, pch = 19)
```



Com podem veure en aquest gràfic, hi ha una lleugera tendència a augmentar la probabilitat de sobreviure quan la edat és menor. Però observem casos de baixa edat amb poca probabilitat de sobreviure i casos de edad avançada amb alta probabilitat.
Si extreiem els individus de baixa edat (primer quartil) amb una probabilitat inferior al 20% de sobreviure, trobem el següent:

```r
p_lage_sur_pre <- passatgers_2[which(passatgers_2$Ppre <= 0.2 & passatgers_2$Age <=
    quantile(passatgers_2$Age)[2]), ]
kable(head(p_lage_sur_pre[3:7]))
```

	Pclass	Name	Sex	Age	SibSp
6	L	Moran, Mr. James	M	21	0
8	L	Palsson, Master. Gosta Leonard	M	2	3
13	L	Saundercock, Mr. William Henry	M	20	0
17	L	Rice, Master. Eugene	M	2	4
37	L	Mamee, Mr. Hanna	M	19	0
38	L	Cann, Mr. Ernest Charles	M	21	0

Podem observar que la majoria dels individus són homes i tots de classe baixa. Com hem dit, el fet de ser home disminueix la probabilitat de sobreviure, com també pertànyer a la tercera classe.

Si extreiem els individus de edat avançada (tercer quartil) amb una probabilitat superior al 70% de sobreviure, trobem el següent:

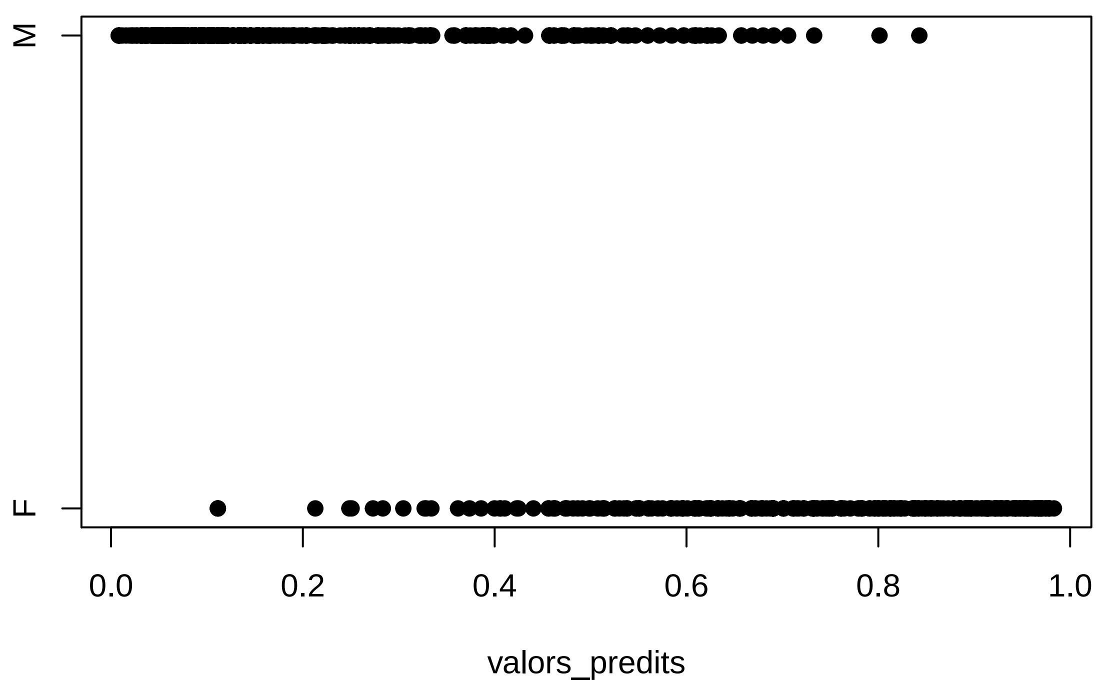
```r
p_hage_sur_pre <- passatgers_2[which(passatgers_2$Ppre >= 0.7 & passatgers_2$Age >=
    quantile(passatgers_2$Age)[4]), ]
kable(head(p_hage_sur_pre[3:7]))
```

	Pclass	Name	Sex	Age	SibSp
2	H	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	F	38	1
12	H	Bonnell, Miss. Elizabeth	F	58	0
53	H	Harper, Mrs. Henry Sleeper (Myna Haxtun)	F	49	1
62	H	Icard, Miss. Amelie	F	38	0
162	M	Watt, Mrs. James (Elizabeth "Bessie" Inglis Milne)	F	40	0
178	H	Isham, Miss. Ann Elizabeth	F	50	0

Podem observar que la majoria dels individus són dones de classe alta.

Per últim, si mostrem un gràfic amb la comparació de la probabilitat de sobreviure amb el sexe dels individus:

```r
stripchart(valors_predits ~ passatgers_2$SexR, pch = 19)
```

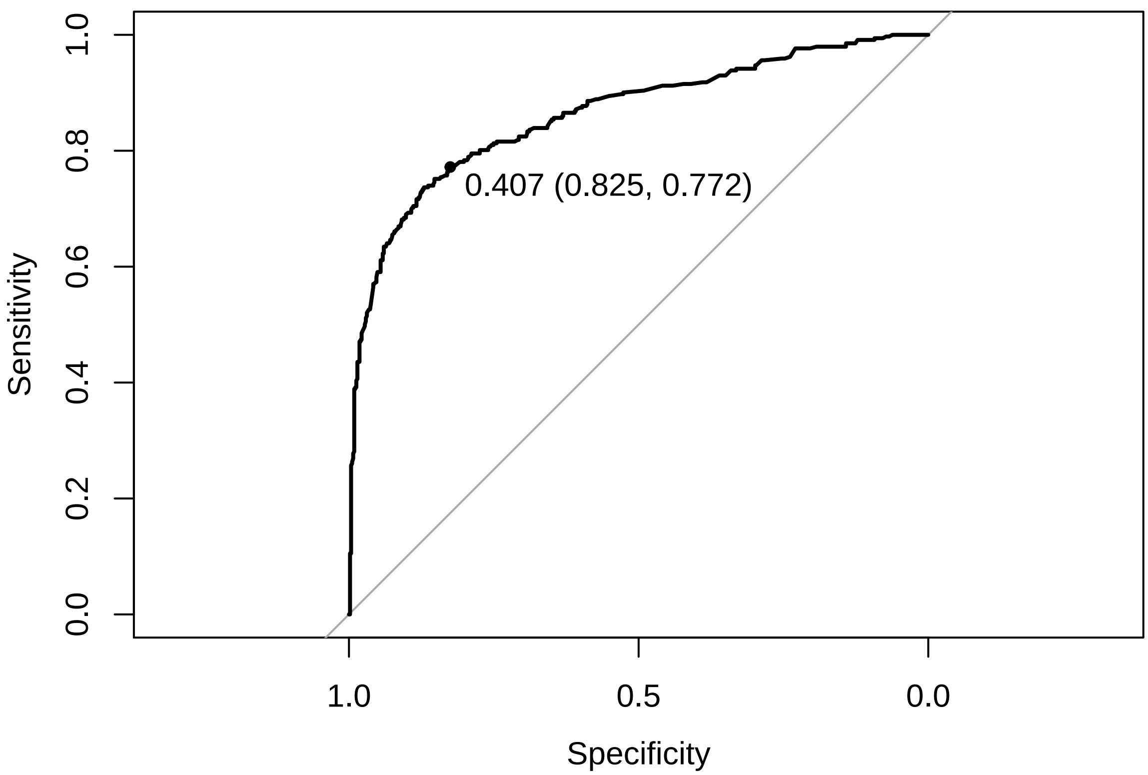


Veiem clarament, que la probabilitat de sobreviure és major en els casos de les dones que en els casos de les dones.

Per tal de mostrar la qualitat del model, podem mostrar la corba ROC associada:

```r
# Calculem la corba ROC
roc <- roc(passatgers$Survived, valors_predits)
# Mostrem la corba calculada en un gràfic
```

```r
plot(roc, print.thres = "best", print.thres.best.method = "closest.topleft")
```



Com podem veure, la forma de la corba s'aproxima molt a la cantonada superiror-esquerra del gràfic, la qual cosa indica que la qualitat del model és alta. Una altra dada que és pot extreure del gràfic anterior és el llindar òptim, és a dir, el llindar de discriminació que maximitza la sensibilitat i la especificitat del model, o amb altres paraules, el que minimitza els falsos positius i falsos negatius.

Si apliquem aquest llindar al model, obtenim la matriu de confusió següent:

```r
clase_predita_2 <- ifelse(valors_predits > 0.407, 1, 0)
# Montem el data set a analitzar
data_2 <- data.frame(obs_2 = passatgers_2$Survived, pre_2 = clase_predita_2)
# Montem la matriu de confusió
matriu_confusio_2 <- table(data_2$obs_2, data_2$pre_2, dnn = c("Observació", "Predicció"))
matriu_confusio_2
##             Predicció
## Observació   0   1
##           0 453  96
##           1  78 264
```